Using Online Classified Ads to Identify the Geographic Footprints of Anonymous, Casual Sex-seeking Individuals

Abstract—This paper describes a method of using Craigslist personal ads to better understand the movement behavior of anonymous, casual sex-seeking individuals within the men-who-have-sex-with-men community. Given recent dramatic increases in HIV and sexually transmitted disease within this community, gaining insight into how sexual networks connect neighborhoods and cities is important for formulating public health interventions. Due to the high degree of similarity exhibited by subsets of Craigslist ads, and the presumption that a set of near-identical ads, when not spam, originate from the same author, we can apply techniques for efficient near-duplicate detection to identify clusters of near-identical ads. By examining each of these clusters and identifying differences in user-supplied location tags, we can then reconstruct an approximation of an anonymous individual’s movement footprint over time, as well as estimate the rate at which ad authors seek sexual encounters. For the state of California, we find that 86% of all encounter requests for a given set occur within a 50 mile area, with only less that 4% of messages reflecting long-distance travel over 250 miles. 60% of all detected clusters reposted ads within 2 weeks of the first detected post. We show that even in the relatively noisy, unstructured data environment of anonymous personal ads, it is still possible to extract meaningful signal and identify useful social network properties for analysis.

I. INTRODUCTION

Casual sexual encounters in the Internet age present a series of unique challenges for public health officials. Sexual encounters arranged online are likely to be geographically broader, anonymous, and greater in number. Bull et al. [1] found via a national survey that 43.3% of MSM (men who have sex with men) and 56.4% non-MSM individuals negotiating sexual encounters via the Internet had traveled 100 or more miles to meet their partner. A meta-analysis estimates that 40% of MSM meet their sex partners online [2] and the Internet has been identified as the largest venue where MSM meet sexual partners [3]. The increased geographic scale of contacts complicates public health efforts since contacts may cut across public health boundaries, weakening traditional public health response tools such as contact tracing and partner notification [4–6]. These developments are of particular importance to intervention strategies within the MSM community, which has experienced a dramatic increase in HIV infections and other sexually transmitted diseases (STDs) since the mid-1990’s. Between 2001-2005 there was a 13% increase in HIV infections [7, 8] and a 10-fold increase in syphilis diagnoses [5, 9] in the U.S. alone.

There are many online venues and social networking sites where an individual can negotiate Internet-mediated sexual encounters, several of which specifically target the MSM community (e.g., Manhunt, Grindr, Adam4Adam, etc.). One of the largest publicly accessible sites is the online classified advertisement website Craigslist: as of July 2011, the 9th most visited website in the US [10]. While Craigslist predominately facilitates the exchange of traditional goods (e.g., cars, musical instruments, furniture) it also supports, via the personals section, a large community of anonymous, casual-sex-seeking individuals, predominantly but not exclusively MSM.

The public nature of these ads present many useful opportunities for surveillance, situated within a known high-risk sexual community. Ad text encodes a variety of risk behaviors with identified links to STDs: safe or unsafe sexual requests, preferences for drug or alcohol use during encounters, and requests for group sex. Many authors also report their HIV status and the preferred HIV status of their partner (serosorting). As separate Craigslist websites exist for individual cites across the United States, this behavioral information is linked to geographic location. All Craigslist ads contain an implicit location by virtue of the site in which ads are placed, but the majority also contain a more specific location reference in the form of a user-supplied location tag of surprisingly high spatial resolution.

STD epidemiology utilizes the concept of core groups to define both the individuals engaging in high-risk behaviors (e.g., repeated sexual encounters, repeat infections, commercial sex work, etc.) and the geographic clustering of outbreaks associated with these groups. Models suggest that core groups are critical to maintaining transmission of disease within a population [11]. Unfortunately, the individuals comprising these groups are difficult to identify and locate, in part because membership within the core varies over time. However, the spatial component or core area does appear
to remain stable over the course of an outbreak, leading to intervention strategies that target risk spaces – the geographic locations linked to sexual encounters [12]. Identifying these risk spaces is usually done through surveys or interviews, which are expensive, slow, and often conducted in response to a detected outbreak, and therefore by definition long past the optimal time for an intervention. Given these constraints, an inexpensive method for efficiently collecting large-scale, location-specific behavioral data would be useful to the public health community.

Additionally, some researchers theorize that STD epidemics have different transmission dynamics depending upon the spatial origin of an outbreak; epidemics originating from within core areas may be more difficult to bring under control than those originating from non-core areas [13]. This suggests that the ability to monitor sexual traffic between geographic clusters may provide additional insight into the progression of an epidemic. While the role of air travel in the spread of diseases such as influenza has been explored by many researchers [14, 15] the impact that local travel has on the dynamics of STDs is less discussed in the literature. Large metropolitan areas, such as Los Angeles, involve some implicit level of local travel, but the degree to which sexual networks connect neighboring cities, or even geographic locations within cities themselves (beyond known clusters of socio-demographic disease correlates), is less understood. Travel across state and county lines all have potential impact on public health department surveillance and intervention strategies, which are ultimately bound by geographic and political boundaries.

We make two observations about anonymous Craigslist personal ads: 1) many users appear to post ads at regular intervals, making little or no textual changes to the body of their message; and 2) changes are often confined to the subject line and contain updated location information. This reposting phenomenon likely stems from the fact that Craigslist, as of March 2012, has no mechanism for users to permanently delete or remove an ad from their account history. Every user’s account management page maintains a log of all posting activity, including ad text, date of post, and ad category. This log, when coupled with Craigslist’s 2-click mechanism for quickly reposting expired ads, creates an incentive for some users to repost ads, either maintaining the exact same content, or – more commonly – making small changes to the text. By examining the set of all ads, removing spam, and identifying these near-duplicate ad clusters among the remaining posts, we can construct sets of ads that likely correspond to a single user’s posting activity over an interval of time. Using the spatial information present in these ad clusters, we can then construct an estimate of the geographic region or footprint where an individual is willing to meet for a casual sexual encounter and calculate associated travel distances within that region. This can provide insight not only into the spatial clustering of behaviors defining core groups, but the degree to which geographic clusters themselves are interconnected.

II. BACKGROUND

A. Craigslist Background

Craigslist is an online classified advertisement service that allows users to post free, anonymous classified advertisements (ads) in a variety of different categories (e.g., items for sale, job offerings, dating personals, etc.). Craigslist is organized around local communities and is structured as a network of sub-websites (sites). Each site contains ads from its primary anchor city or state, as well as from smaller surrounding communities. Within the U.S., some states consist of a single site for the entire state while other states may encompass between 4 and 28 sites. Each site contains a set of standardized categories and all ads are publicly accessible via RSS (Really Simple Syndication) feeds.

From July 1, 2009 until July 1, 2011, Craigslist RSS feeds were collected once a day for 8 personal ad categories in 414 sites across the United States. This 2-year collection of RSS feeds forms the document collection or corpus examined in this paper and consists of over 67 million personal ads. All Craigslist ads contain one or more encounter tags characterizing the type of individual the author is seeking, which are used to restrict our analysis to just ads targeting the MSM community. Personal categories containing MSM-related ads are as follows: m4m (men seeking men); cas (casual encounters); msr (miscellaneous romance); and sip (strictly platonic). Filtering out unrelated encounter tags results in a final ad set size of over 33 million MSM-specific ads.

B. Craigslist & STD Surveillance

The links between Craigslist and STDs has been explored by a number of researchers, with some preliminary research suggesting that the entry of Craigslist into the market can itself be linked to an increase in HIV/AIDS and syphilis rates in the U.S. [16]. Moskowitz and Seal explored the connection between ad posting frequency and MSM health outcomes, finding that men who frequently posted ads resulting in sexual encounters reported more negative health behaviors and STD rates [17]. Grov examined MSM ads posted in New York City, manually developing guidelines for annotating risk behaviors in the text of ads [18]. In our previous work, we examined HIV and behavioral surveillance applications by building phrasal and keyword-based lexicons for semi-supervised ad annotation and geolocation [19, 20]. We found that in California, self-reported HIV rates were highly correlated with county-level HIV/AIDS prevalence, and that monitoring terminology associated with high-risk behaviors (e.g., unprotected sex requests, drug use during sex, etc.) can, at the ecological level, be used to predict yearly, county-level syphilis incidence.

C. Near Duplicate Detection

The near-duplicate detection problem is the task of efficiently identifying textual content or documents in a collection that, using a pairwise measurement of similarity, fall within a similarity threshold value in the range [0, 1]. Detecting near-duplicates has many applications in web-crawling [21], plagiarism detection [22], and document clustering [23]. Our
task is similar to that of plagiarism detection, in that we wish to detect repeated subsequences of text across many different Craigslist ads. The idea that various word/phrase choices, text formatting (e.g., consistently capitalizing words for dramatic effect), and other author-specific linguistic features or *stylometry* can be used to identify individuals in online contexts was explored in [24]. We do not consider any of these more advanced approaches to identifying online identity in anonymous settings, but instead focus on identifying nearly-identical ads using \( n \)-gram document features (i.e., sequences of \( n \) consecutive words).

Since exactly solving common similarity functions such as the Jaccard index are computationally expensive and therefore intractable for large document collections [25], a number of locality-sensitive hashing techniques have been developed for efficiently identifying approximate sets with similar features. Charikar’s simhash [26] is a probabilistic dimension reduction technique by which a set of input features is hashed into a \( f \)-bit long fingerprint such that similar input items hash to similar values with very high probability.

### III. METHODS

#### A. Craigslist Ad Preprocessing

1) Geocoding Ads: Every ad contains a user-supplied location tag corresponding the location or region where the author wishes to negotiate a sexual encounter. These tags can be thought of as analogous to “check-ins” found in Location-based Networks (LBNs) such as FourSquare, but with a highly variable spatial resolution and encoding a number of superordinate-subordinate relationships. For example, some tags correspond to metropolitan-sized areas (e.g., “L.A.”) while others are venue-specific (e.g., “Dolores Park”, “Marriott Marquis”, etc.) and capture the spatial resolution typically seen in LBNs.

To assign geocode coordinates to each ad, respective of this spatial variability, we first built a site-wise frequency distribution of all Craigslist location tags originating from the state of California. Using 2009 & 2010 geographic boundary files [27, 28], we then created a hierarchical named entity gazetteer of all counties, cities, roads, airports and other major landmarks for the state. The resulting set of **named entities** was then compared to the set of Craigslist location tags using Jaro-Winkler distance as a measure of string similarity [29]. Jaro-Winkler distance is specifically designed for matching proper names, given the observation that spelling variations occur more frequently within the ending letters of proper names than at the beginning. Tag-to-entity mappings were assigned whenever a location tag matched a named entity within a distance threshold of 0.95. Any location tag that could not be identified using our gazetteer (typically tags corresponding to business names such as Hilton or Marriott) was geocoded using the Google Maps API. Ads that could not be geocoded using either approach were removed from our analysis.

To validate that tag coordinates were geographically plausible, for each site we defined a *Craigslist coverage area*, representing the geographic region serviced by the site in question. These coverage areas are defined as the convex hull bounding the metropolitan statistical area of the site’s primary city (e.g., Los Angeles) or, for areas that do not directly correspond to cities (e.g., Gold Country, CA), the set of cities and towns contained within that region. The list of such cities is extracted from the Wikipedia page for each region, which Craigslist provides a link to on each site’s homepage. Any coordinates not contained inside or within 20 miles of this boundary were flagged for manual validation; the resulting set of tags was then manually inspected and any spurious tags removed.

#### B. Identifying Near-Duplicate Ads

1) Fingerprint Generation: To calibrate our selection of document features and evaluate the detection accuracy of our parameter choices, we conducted several preliminary experiments. All Craigslist ads from the San Francisco Bay Area site were tokenized into case-sensitive unigrams and bigrams (i.e., single words and adjacent word pairs) and a 64-bit simhash fingerprint created for each resulting document term vector. We then calculated the all-pairs Hamming distance [30], \( k \), for each ad. 50 pairs of hashes at Hamming distances \( k = [1...10] \) were randomly selected from each \( n \)-gram fingerprint set. A human annotator then evaluated each pair as a true or false positive near-duplicate match. The guidelines for making decisions were as follows: pairs were flagged as false positives when they contained little to no sentence or phrasal commonalities; true positives required that ads share substantial matching textual content, orthographic features (e.g., capitalization, slight differences in sentence order, and/or minor textual changes). Based on these results (see Figure 1) we choose a Hamming distance of 6-bits for our detection threshold.

![Figure 1](image.png)

Fig. 1. True positive rates for hash fingerprint feature selection (using unigram and bigram feature sets) and evaluated at Hamming distances \( k = [1...10] \). Note that unigram features perform poorly overall, while bigrams perform well at distances \( 1 \leq k \leq 7 \).

To efficiently compute the pairwise Hamming distance of all 33M ads, we utilized the method described in Manku et al. [21]. The key observation to their approach is that the top \( n \)-bits of each fingerprint, when sorted, function approximately as a counter, allowing queries to quickly reduce the search space necessary for finding hashes within some distance threshold. All 64-bit fingerprints are split into 4 tables of 16-bit blocks, with each table’s blocks permuted such that each has

---
a different leading block of bits. Each ad’s fingerprint is then queried against these 4 tables and all fingerprints within \( k=6 \) bits of the query are returned and greedily assigned to a cluster. Ads with no matches are considered singleton messages. Ads that are previously assigned to a cluster are not queried.

2) Removing Spurious Ad Clusters: The end result of the query process consists of the set of all singleton messages and \( n \) matched ad clusters. Because these ad clusters contain both a level of spurious matches (as a result of the fingerprinting process) and a degree of multi-post spam messages, additional filtering is required to identify clusters corresponding to the behavior of real users over some interval of time.

First, to filter out spurious hash matches, for each cluster we calculated the pairwise inter-cluster Jaccard index for the corresponding ad term vectors. Any cluster in which the median similarity value falls below a threshold \( r \) (we used 0.60) is discarded (see Figure 2 for the distribution of inter-cluster similarities). Likewise any single document contained within a cluster that falls below this threshold is also removed.

The remaining clusters are further processed to address ads reposted multiple times within a small time window to the same Craigslist site, which seem to typically result from users making minor changes or corrections to the text content of an ad and reposting. Ads within a cluster that are posted within 4 hours of each other are collapsed into 1 posting instance. Clusters consisting solely of near-duplicates within the same 24-hour interval are also considered singleton ads.

3) Removing Spam: Identifying near-duplicate ads on Craigslist also identifies spam ad clusters. These spam clusters can consist of web advertisements, chain letters, or requests unrelated to negotiating a direct-contact encounter (e.g., ads asking to chat over webcams, posting nationally to trade erotic pictures, etc.). All of these spam types exhibit wide geographic dispersion, with multiple ads posted within a narrow window of time to numerous, geographically distant Craigslist sites, making it highly unlikely they correspond to realistic travel intentions. While Craigslist maintains internal mechanisms for detecting both similar content posted across multiple sites and duplicate ads posted within the same 48-hour window, manual inspection of detected clusters reveals that spam still exists in our corpus.

We removed any cluster containing references to external URLs (e.g., guyflings.com, caliguys.com, etc.) and any cluster using inline HTML markup to obfuscate text for bypassing Craigslist’s internal spam filters (e.g., “Go<tr>o<tr>gle<tr> fo<tr>r i<tr>r”). To filter out any remaining spam clusters, we identified several non-lexical features of spam clusters and used a machine learning approach to classify them. Certain Craigslist spam ad clusters exhibit features similar to those observed in botnets [31], where spam can be characterized, in part, by rapid bursts of posting activity across a widely dispersed geographic area. Since our corpus covers the entire United States, we can extract a geographic distribution for all clusters and identify clusters with large, implausible geographic and temporal footprints. All feature are calculated as a vector of values, aggregated into 7-day intervals, over the duration (lifespan) of each ad cluster.

Cluster Features

- **Maximum Weekly Burst Rate:** the maximum number of ads posted within any single 7-day interval.
- **Mean and Maximum Weekly Geographic Dispersion:** for all ads ordered in time, we calculate the sum of all hop distances within a single 7-day interval, and record the mean and maximum value of the resulting vector.
- **Geotemporal Stability:** for the duration of the cluster, the percentage of weeks in which there was 0 dispersion, i.e., either no ad was posted or the ad contained no change in location.

The entire ad cluster corpus was split by ad cluster size (i.e., the number of ads comprising a single cluster) into 10 bins and clusters randomly selected from each bin. In total, 525 ad clusters were manually inspected and flagged as spam or valid. Guidelines for flagging spam involved ordering all posts within a cluster by time and examining the hop distances between each posted ad location; clusters involving multiple long distance hops within a short window of time were flagged as spam. We used WEKA [32] to test a number of classifiers, of which an AdaBoostM1 classifier evaluated with a DecisionStump as the base classifier performed best, with a weighted average true positive score of 93.5% and ROC area value of 0.96.

4) Constructing Footprint Travel Graphs: We built 2 directed travel graphs for the state of California, representing the change in location tags within clusters over time. The first graph contains all entity-mapped location tags, while the second aggregates all tags into their county of origin. Edges are created using pairs of ads within clusters, where all ads are ordered in time and each edge represents a location tag change between each ad comprising the pair. For example, if the first ad in a cluster contained the location tag “glendale” (a city in Los Angeles) and the next detected ad within that same cluster contained the location tag “long beach” (a city also in Los Angeles, but 30 miles south) a directed edge (Glendale, Long
(Beach) is created connecting both locations. Edge weights were assigned in units of w ads, with w taking on fractional values in circumstances where a single location tag contained multiple locations (e.g., “burbank/north hollywood/glendale”). In these cases, edges were created for each pair of tags across ads. For the county graph, edges are only created when changes to location tags cross county lines. Given these edges, we then calculated the haversine (or great circle) distance for all paths to derive estimates for the “roaming” geographic distance covered by each cluster’s footprint. Table II contains graph summary statistics for both footprint graphs.

IV. Results

A. Near-Duplicate Identification Summary

<table>
<thead>
<tr>
<th>Table I</th>
<th>Ad Cluster Sizes (U.S. and California)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S.</td>
<td>California</td>
</tr>
<tr>
<td>Spam</td>
<td>Cluster N</td>
</tr>
<tr>
<td>65,825</td>
<td>165,208</td>
</tr>
<tr>
<td>Collapsed to Singleton</td>
<td>36,202</td>
</tr>
<tr>
<td>Valid Clusters</td>
<td>1,130,886</td>
</tr>
<tr>
<td>≥ 4 ads per Cluster</td>
<td>249,185</td>
</tr>
<tr>
<td>Cross-site Clusters</td>
<td>10,516</td>
</tr>
<tr>
<td>Cross-state Clusters</td>
<td>7,218</td>
</tr>
</tbody>
</table>

Table I contains summary statistics for the final filtered set of ad clusters for both the entire United States and the state of California. These remaining filtered clusters represent likely (though still anonymous) posting behavior of a single Craigslist user, where each ad can now be viewed as a check-in to a given location at some point in time. These check-ins define the geographic region within which an ad’s author is willing to travel for a casual sexual encounter. The textual changes authors make to their ads over time reveal useful information and can be used to gain insight into changes in behavior. By examining the changes in each cluster’s set of user-supplied location tags, for example, we can calculate how the geographic region changes over time and derive an estimate of how far that author is willing to travel for an encounter. Users also update biographic details, their age for example, which can provide a weak form of validation. Over the entire set of ads reporting age (14% of all clusters in California), the mean age change was 0.25 years (S.D. 1.4 years). Of the 1,410 ad clusters spanning longer than one year (0.6% of the total clusters), 67% of the clusters contain changes incrementing their reported age by 1 or 2 years, with a mean change of 2.3 years (S.D. 2.6) over the 2-year corpus.

B. Post Rate and Interval Duration Distributions

Figure 3 shows the empirical cumulative distribution function of the average interval between ad reposts (calculated for all clusters containing more than 4 ads) which can be viewed as an estimate of the periodicity of individual authors’ Craigslist usage patterns. 50% of clusters repost an ad within 2 weeks of the initial detected post and 50% of all clusters last 30 days or less. Overall we see that our system detects only a very small subset of authors who post frequently for long intervals of time; the majority of detected ads fall within the 3-30 day interval. This is likely a result of the fact that authors, if they continue using Craigslist, eventually change their ad textual content such that the similarity to previous ads falls outside our detection threshold, or they write entirely new ads.

C. Footprint Distances and Location Entropy

Figure 4 shows the distribution of per-hop roaming distances for all detected California ad clusters. The majority of all cluster paths involve footprint distances of no more than 50 miles, with a small percentage (4%) indicating long-distance travel. The most common of these paths was between Los Angeles and San Francisco, as well as paths to outlining mountainous counties such as El Dorado and Placer County (the counties adjoining Lake Tahoe, a well-known vacation destination).

Figure 5 shows the hourly average Shannon entropy of ad location tags and average post counts for the duration of our corpus. A higher entropy value corresponds to more observed variability in location tags for that hour. All location tag counts are binned by U.S. state of origin. Note how entropy periodicity mirrors posting frequency, with two distinct morning and evening posting peaks. While Tuesday-Friday involve more evening posting on average, the corresponding entropy of location tags is essentially equivalent between peaks. Weekend (Saturday and Sunday) mornings show a slightly higher entropy overall, revealing that those up in the early morning hours express more willingness to travel than
Fig. 4. Empirical Cumulative Distribution Function (top) and histogram (bottom) of hop distances, calculated as the haversine distance (in miles) between location tags of successive ads within a given ad cluster. 86% of all hop distances involve distances under 50 miles while 4% involve distances over 250 miles. Note as the haversine distance is the minimal great-circle distance between 2 points on a sphere, these values underestimate the actual number of miles traveled.

Fig. 5. (Top) Average hourly entropy of location tags counts, binned by U.S. state, and (bottom) average posting counts by hour for all U.S. clusters. Black dashed lines divide days and grey dotted line corresponds to noon. Observe how the periodicity of tag entropy essentially corresponds to posting frequency, with slightly more location tag variability (i.e., more willingness to travel) observed Friday and Saturday evenings and Saturday and Sunday mornings.

Fig. 6. Visualization of the Craigslist sfbay subgraph (V=591, E=7916), showing all edges with weights > 25. Node color represents outdegree and edge color represents weight (blue=low and red=high). Observe how geographic clustering is clearly visible with respect to the spatial distribution of city locations.

Table II

<table>
<thead>
<tr>
<th></th>
<th>California Graphs</th>
<th>All Location Tags</th>
<th>County-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>4,464</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>Edges</td>
<td>43,023</td>
<td>792</td>
<td></td>
</tr>
<tr>
<td>Avg. Clustering Coefficient</td>
<td>0.022</td>
<td>0.566</td>
<td></td>
</tr>
<tr>
<td>Avg. Degree</td>
<td>0.301</td>
<td>14.4</td>
<td></td>
</tr>
<tr>
<td>Diameter</td>
<td>15</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Avg. Path Length</td>
<td>5.7</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Min Edge Weight</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Max Edge Weight</td>
<td>154</td>
<td>445,978</td>
<td></td>
</tr>
</tbody>
</table>

D. Travel Networks

Figure 6 shows the San Francisco Bay Area as a directed graph of changes in location tags, which can be viewed as a measure of flow between locations, in terms of an individual’s willingness to travel to that location for a sexual encounter. Observe how the layout respects the actual spatial distribution of cities in the Bay Area and how most flow occurs between downtown San Francisco and San Jose – the result of both legitimate travel between these two locations and the semantic ambiguity of the tag “downtown” which may also refer to downtown San Jose. Figure 7 shows the county-level, directed graph of travel between counties in California. Note how some counties are characterized by asymmetric inward/outward flow. This can be partly explained by the bias inherent in how people report location tags – authors are more likely to report tags about locations they intend to travel to – as well as the asymmetric desirability of certain long-distance travel locations (i.e., vacation spots).

V. Conclusion

We have shown that a near-duplicate identification system can successfully and efficiently cluster Craigslist ads with similar content. Moreover, even in the noisy, unstructured data early morning/late night posts made during the work week.
environment of anonymous personal ads, we show that it is still possible to extract meaningful signal. By leveraging the information contained within these clusters, we are able to construct graph representations of the geographic footprints of sexual encounter requests at varying levels of spatial resolution. We also report data capturing an author’s willingness to travel within these footprints in miles, the first such results that did not originate from survey data. We find that the majority of encounter requests occur within a 50-mile area around a given location, with a small percentage of travel involving distances of 100 or more miles.

A. Limitations

There are a number of limitations to this work. First, Craigslist ads only reflect an intended sexual encounter, not necessarily that an encounter occurred. However, ads still provide insight into the behavior and sexual norms of a known high-risk community, making them a useful surveillance resource. Secondly, the relationship between the location of a sexual encounter and the home locations of the participants is unknown, as an author may prefer to travel for this purpose, with concomitant implications for formulating interventions. However, since there has been some success with intervention strategies targeting risk-spaces (e.g., bars and clubs, places of prostitution, etc.) [33, 34], identifying locations with high rates of casual encounters is still of great practical value to the public health community. Third, there is no way of directly validating that our detected ad clusters are in fact written by the same individual. While it is possible that some ad clusters consist of text copied from other ads (resulting in multiple individuals being collapsed into a single cluster) we feel that various social incentives exist within Craigslist to encourage people to honestly represent their sexual interests. In addition to self-policing mechanisms for flagging spam ads, many authors will post ads warning other community members about authors that consistently lie or misrepresent themselves. This fact and the practical Craigslist user-interface incentives for reposting or copying an author’s own text suggest that many clusters legitimately originate from the same author. Finally, since our method relies on calculating the maximum distances between centroids of locations, our distance values provide only approximate estimates of how far an ad author would actually have to travel. Yet, despite these limitations, we feel that understanding the relationship between cities and the movement footprints of casual-sex seeking individuals at a scale and temporal resolution hereto unexplored can provide meaningful new insights into the spread of STDs.

B. Future Work

Feature selection for constructing fingerprints was limited to simple un-weighted bigram document vectors. In the future, a more thorough exploration of feature selection (e.g., stemming, phrase identification, etc.) may result in better near-duplicate detection, increasing both the number of clusters identified and the size of detected clusters. Manual inspection of ad cluster content reveals that a number of clusters themselves are likely written by the same author, as identified by certain unusual, low frequency phrase choices and other orthographic features (e.g., consistently capitalizing the same words or phrases across many ads). This suggests that fingerprinting messages with additional orthographic-level features may provide additional ways of identifying clusters.

In the terminology of information retrieval, our current system has high precision, but suffers from low recall. The clusters we identify are likely genuine, but as a result of our chosen feature set we fail to identify the majority of clusters that exist within our corpus. Currently our framework can detect long-distance travel (such as an author traveling from New York, NY to Los Angeles, CA), but such clusters are found infrequently; improving recall may lead to the ability to detect this phenomena more often.

The temporal aspect of footprint size also remains unexplored. The degree to which a footprint remains stable or changes over time may provide information regarding an author’s success in locating partners within a region, or their cognizance of outbreak conditions within certain locations.

REFERENCES

